ANNALES
UNIVERSITATIS MARIAE CURIE-SKŁODOWSKA
LUBLIN—POLONIA
VOL. XVIII, 9 SECTIO H 1984

Zakład Zastosowań Matematyki
Wydział Ekonomiczny UMCS

Mieczysław SOBCZYK

Wielowymiarowa analiza statystyczna

Многоизмерительный статистический анализ

Multi-dimensional Statistical Analysis

W każdym badaniu statystycznym można wyodrębnić trzy wzajemnie ze sobą powiązane etapy, a mianowicie: gromadzenie informacji, ich porządkowanie i klasyfikację oraz modelowanie. Ostateczny wynik badania statystycznego zależny jest od prawidłowego przeprowadzenia prac w poszczególnych etapach, przy czym istotną rolę odgrywa tu ich kolejność. Oznacza to, że wartość poznawcza otrzymanych modeli kształtowania się zjawisk masowych zależy po pierwsze od rzetelności, obfitości i trafnego doboru informacji statystycznych, po drugie zaś — od właściwej klasyfikacji zebranych danych. Informacje statystyczne niezbędne w badaniu czerpie statystyk z ogólnie dostępnej sprawozdawczości i ewidencji gospodarczej. Z reguły — poza szczególnego typu badaniami ankietowymi — nie ma on wpływu na ich jakość. Dlatego też tym większej wagi nabiera problem właściwego porządkowania i klasyfikacji materiału statystycznego.

Klasyfikacja — w najbardziej ogólnym ujęciu — jest działem metodologii ogólnej stanowiącej jedno z podstawowych narzędzi badania rzeczywistości.[1] Termin klasyfikacja odnosi się także do samej czynności podziału elementów zbioru na podzbiory (człony klasyfikacji) oraz do końcowego wyniku tej czynności, tj. do otrzymanych z podziału zbioru podzbiorów. W ujęciu teoriomnogościowym klasyfikacja stanowi niepustą rodzinę podzbiorów S_i $(i = 1, 2, ..., k)$ pewnego zbioru, spełniającą dwa warunki: rozłączności $(S_i \cap S_j = \emptyset,\quad i \neq j,\quad i, j = 1, 2, ..., k)$ oraz zu-

[1] T. Wójcik: *Zarys teorii klasyfikacji.* Warszawa 1965.

pełności ($\bigcup_{i=1}^{k} = \Omega$). Jako synonimów terminu klasyfikacja używa się również takich określeń, jak porządkowanie, dyskryminacja, delimitacja, taksonomia.[2] W przypadku wykorzystywania w procedurze klasyfikacji metod ilościowych używa się często terminów: taksonomia numeryczna, taksonometria, taksometria.

W pracach cybernetycznych zamiast powyższych terminów zwykle używa się określeń: teoria rozpoznawania obrazów bądź też teoria układów rozpoznających, uczących się. Teoria rozpoznawania obrazów wykorzystywana jest przy porządkowaniu dużych zbiorów informacji statystycznych bądź też przy wyodrębnianiu pewnych podzbiorów.

Obrazem nazywamy zbiór realnie lub potencjalnie istniejących obiektów należących do tej samej klasy podobieństwa, charakteryzujących się pewnymi wspólnymi własnościami (cechami). Obrazem jest np. zbiór cech charakteryzujących równowagę rynkową, czy też zbiór osób, które nabyły samochód w pewnym okresie, czy też zbiór przedsiębiorstw wykonujących plany.

Elementy składowe obrazu są obiektami. Proces przyporządkowania nowych, dotychczas nie rozpatrywanych obiektów do danej klasy podobieństwa (obrazu) nazywamy rozpoznawaniem obrazów. Przyporządkowanie obiektów do poszczególnych obrazów odbywa się w drodze porównywania właściwości klasyfikowanego obiektu z właściwościami obiektów należących do pewnej, ustalonej już klasy (obrazu). Zbiór wszystkich obiektów będących przedmiotem klasyfikacji nazywamy przestrzenią prób. Natomiast zbiór wszystkich obrazów (klas podobieństwa) opisanych na danej przestrzeni prób określamy mianem alfabetu klas (obrazów).

Wyjściowym punktem klasyfikacji jest określenie jej przedmiotu i przestrzeni. Przedmiotem klasyfikacji jest zbiór obiektów, którymi mogą być jednostki przestrzenne (np. województwa, gminy), przedsiębiorstwa przemysłowe, handlowe itp. Ogólnie można stwierdzić, że przedmiotem klasyfikacji jest zbiór indywiduów (obiektów) dowolnego rodzaju. Zbiór ten oznaczamy, symbolem Ω, a elementy tego zbioru — symbolem ω_i. Tak więc $\omega_i \in \Omega$, $i = 1, 2, \ldots, k$, gdzie k oznacza liczbę obiektów podlegających badaniu. W zależności od przedmiotu klasyfikacji można wyróżnić klasyfikację przeprowadzoną drogą podziału logicznego oraz przez grupowanie.[3]

---

[2] Por. m.in. W. Bukietyński, Z. Hellwig, K. Królik, A. Smoluk: *Uwagi o dyskryminacji zbiorów skończonych*. Prace Naukowe WSE Wrocław, 1969 nr 21; B. Kopociński: *Dyskryminacja za pomocą dendrytów*. „Zastosowania Matematyki" 1960, nr 3.

[3] Z. Chojnicki, T. Czyż: *Metody taksonomii numerycznej w regionalizacji geograficznej*. Warszawa 1973, s. 8.

Klasyfikacja przez podział logiczny (zwana też klasyfikacją dedukcyjną lub „od góry") dokonywana jest w oparciu o pewne kryterium zapewniające poprawny podział logiczny, tj. realizujące warunki rozłączności i zupełności. Kryterium to jest definiowane z góry. Najprostszym przykładem takiej klasyfikacji jest podział dychotomiczny (dwudzielny). Wyjściowy zbiór obiektów Ω dzielony jest tu na dwa podzbiory (człony klasyfikacji): jeden z nich obejmuje obiekty posiadające pewną cechę, drugi natomiast — obiekty nie posiadające jej.

Klasyfikacja przez grupowanie (zwana klasyfikacją indukcyjną lub „od dołu") odbywa się w drodze grupowania obiektów tworzących zbiór Ω na podstawie ich podobieństwa. Procedura grupowania wymaga tu ustalenia kryteriów dodawania elementów zbioru Ω. W ten sposób np. $\{x_1\} \cup \{x_2\} = S_1$, a $\{x_3\} \cup \{x_4\} \cup \{x_5\} = S_2$. Wynika stąd wniosek, że w klasyfikacji przez grupowanie zachodzi konieczność enumeracji zbioru Ω, podczas gdy w podziale logicznym jest on definiowany. Fakt ten w przypadku podziału logicznego może prowadzić do otrzymywania klas (podzbiorów) pustych, co nie jest możliwe w procedurze grupowania. Dodać przy tym należy, że w ramach podzbiorów (członów klasyfikacji) uzyskanych w pierwszym etapie, można dokonywać dalszej klasyfikacji, co prowadzi do klasyfikacji wielostopniowej. Wynik klasyfikacji wielostopniowej otrzymany w drodze podziału logicznego zależy nie tylko od wybranych cech stanowiących kryterium klasyfikacji, ale również od kolejności, w jakiej cechy te stanowią podstawę podziału. Stąd też istotna jest tu znajomość struktury zbioru będącego przedmiotem klasyfikacji. W klasyfikacji przez grupowanie nie ma potrzeby ustalania hierarchicznego porządku cech.

Przestrzeń klasyfikacji wyznaczona jest przez zbiór własności (cech) opisujących elementy zbioru Ω podlegające klasyfikacji. Elementy przestrzeni klasyfikacji (cechy) powinny być tak dobrane, by spełniały określone wymogi natury zarówno formalnej, jak i merytorycznej. Nie jest przy tym możliwe wskazanie jednej generalnej recepty na właściwy dobór cech, gdyż zależy to od charakteru, przedmiotu i celu konkretnego badania. Niemniej jednak przyjmuje się, że zbiór cech jest wysoce diagnostyczny, jeżeli jego poszczególne elementy spełniają następujące warunki: [4]

1) ujmują najbardziej istotne właściwości analizowanych zjawisk,

2) są proste, jasno sprecyzowane i logicznie ze sobą powiązane,

3) są bezpośrednio lub pośrednio mierzalne oraz dadzą się wyrazić za pomocą wielkości stosunkowych lub absolutnych,

---

[4] J. Fierich: *Próba zastosowania metod taksonomicznych do rejonizacji systemów rolniczych w woj. krakowskim.* „Myśl Gospodarcza" 1957, nr 1.

4) charakteryzują się wysoką zmiennością w przekroju klasyfikowanych obiektów,

5) są nieskorelowane, ale jednocześnie wykazują dużą zależność z cechami nie uwzględnionymi bezpośrednio w procedurze klasyfikacji.

Obiekty, będące elementami składowymi zbioru Ω, charakteryzowane są zwykle przez większą liczbę cech. Stąd też każdy obiekt $\omega_i \in \Omega$ może być rozumiany jako wektor:

$$\omega_i = (x_{i1}, x_{i2}, \dots, x_{in}) \tag{1}$$

gdzie x_{ij} oznacza j-tą składową tego wektora, czyli wartość j-tej cechy posiadaną przez obiekt ω_i. W interpretacji geometrycznej poszczególne obiekty są punktami w przestrzeni n-wymiarowej (stąd nazwa wielowymiarowa analiza statystyczna). Punkty te należy rozdzielić na pewną (ustaloną z góry lub też nie) ilość rozłącznych i wyczerpujących skupisk homogenicznych w sobie i heterogenicznych pomiędzy sobą. Oznacza to, że poszczególne skupiska (klasy, człony klasyfikacji) powinny zawierać punkty położone blisko siebie w sensie ustalonej a priori metryki odległości, a równocześnie znacznie oddalone od punktów należących do pozostałych skupisk.

Zbiór danych wyjściowych stanowiących podstawę klasyfikacji tworzy macierz obserwacji o postaci:

$$X = \begin{vmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ - & - & - & - \\ x_{k1} & x_{k2} & \dots & x_{kn} \end{vmatrix} \tag{2}$$

gdzie: k — liczba obiektów,
n — liczba cech,
x_{ij} — wartość j-tej cechy w i-tym obiekcie.

W macierzy (2) dla każdego obiektu przeznaczony jest jeden wiersz a dla każdej cechy — jedna kolumna.

Cechy opisują różne właściwości badanych obiektów i wyrażane są w różnych miarach. Stąd też nie należy w dalszych obliczeniach posługiwać się bezwzględnymi wartościami cech, lecz ich miarami relatywnymi. Przekształcenia rzeczywistych wartości cech w wielkości relatywne dokonuje się w drodze ich standaryzacji (normalizacji). Normalizacji cech najczęściej dokonuje się następująco:

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j} \tag{3}$$

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{\bar{x}_j} \tag{4}$$

$$x'_{ij} = \frac{x_{ij}}{\bar{x}_j} \tag{5}$$

$$x'_{ij} = \frac{x_{ij}}{S_j} \tag{6}$$

gdzie: i — obiekt badania; $i = 1, 2, ..., k$,
j — cecha; $j = 1, 2, ..., n$,
x_{ij} — rzeczywista wartość j-tej cechy dla i-tego obiektu,
x'_{ij} — standaryzowana wartość j-tej cechy dla i-tego obiektu,
$\bar{x}_j$ — średnia wartość j-tej cechy,
S_j — odchylenia standardowe j-tej cechy.

Każda standaryzacja zmniejsza wariancję cech, a tym samym zniekształca wyniki badań. Jednakże w wielowymiarowej analizie statystycznej standaryzacja jest zabiegiem koniecznym.

W problemie klasyfikacji, oprócz określenia przedmiotu i przestrzeni, niezbędny jest wybór odpowiedniego kryterium klasyfikacji. Kryteria klasyfikacji są funkcjami podobieństwa przyporządkowującymi każdej parze elementów $\omega_i, \omega_j \in \Omega$ miarę ich wzajemnego podobieństwa (niepodobieństwa). Miarami tymi są najczęściej odległości lub współczynniki podobieństwa. Wybór funkcji podobieństwa uzależniony jest przede wszystkim od charakteru cech uwzględnionych w badaniu. I tak w odniesieniu do zmiennych (cech) ciągłych z reguły stosuje się odpowiednio określone odległości. Natomiast w przypadku zmiennych binarnych stosuje się głównie współczynniki podobieństwa.

W badaniach empirycznych najczęściej wykorzystywane są odległości Euklidesa. Zbiór obiektów Ω traktowany jest jako przestrzeń metryczna, co umożliwia każdej parze jego elementów ω_i, ω_j przyporządkować dokładnie jedną nieujemną liczbę rzeczywistą $d(\omega_i, \omega_j)$ spełniającą następujące aksjomaty:

1) aksjomat tożsamości ($d_{ij} = 0$ wtedy i tylko wtedy gdy $i = j$),
2) aksjomat symetrii ($d_{ij} = d_{ji}$),
3) aksjomat trójkąta ($d_{is} + d_{sl} \geqslant d_{ie}$)

Odwzorowanie d określone na zbiorze wszystkich par elementów przestrzeni Ω nazywamy metrykę przestrzeni metrycznej Ω. Elementy tej przestrzeni nazywamy punktami, a wartość odwzorowania $d(\omega_i, \omega_j)$, czyli wartość metryki, nazywamy odległością punktu ω_i od punktu ω_j.

Odległości euklidesowe najczęściej obliczane są jako:

1) przeciętne bezwzględnych różnic wartości cech:

$$d_{rs} = \frac{1}{n} \sum_{j=1}^{n} \left| x'_{rj} - x'_{sj} \right| \tag{7}$$

gdzie: d_{rs} — odległość między obiektem r-tym oraz s-tym dla $r \neq s = 1, 2, ..., k$
x'_{rj} — standaryzowana wartość j-tej cechy w r-tym obiekcie $(j = 1, 2, ..., n)$
x'_{sj} — standaryzowana wartość j-tej cechy w s-tym obiekcie,
n — liczba uwzględnionych cech,

2) jako pierwiastek z przeciętnej kwadratów różnic wartości zmiennych (cech):

$$d_{rs} = \left[\frac{1}{n} \sum_{j=1}^{n} (x'_{rj} - x'_{sj})^2 \right]^{1/2} \tag{8}$$

3) jako sumę bezwzględnych różnic wartości zmiennych:

$$d_{rs} = \sum_{j=1}^{n} \left| x'_{rj} - x'_{sj} \right| \tag{9}$$

4) jako pierwiastek z sumy kwadratów różnic wartości zmiennych:

$$d_{rs} = \left[\sum_{j=1}^{n} (x'_{jr} - x'_{sj})^2 \right]^{1/2} \tag{10}$$

Po obliczeniu odległości każdego obiektu od wszystkich pozostałych w danym zbiorze Ω otrzymujemy macierz odległości o postaci:

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1k} \\ d_{21} & d_{22} & \dots & d_{2k} \\ - & - & - & - \\ d_{k1} & d_{n2} & \dots & d_{kk} \end{bmatrix} \tag{11}$$

Odległości zawarte w macierzy D zostały określone w przestrzeni, której wymiary wyznacza liczba uwzględnionych zmiennych (cech). Macierz D stanowi podstawę dalszych operacji statystycznych zmierzających do uzyskania jednorodnych podzbiorów (wynik klasyfikacji).

Zwrócić należy uwagę na fakt, że przedmiotem procedury klasyfikacyjnej mogą być zarówno obiekty, jak i cechy. W pierwszym przypadku odległości obliczane są między punktami identyfikowanymi przez wiersze wyjściowej macierzy obserwacji (zwykle zestandaryzowanej), w drugim zaś — pomiędzy punktami, którym odpowiadają kolejne kolumny tej macierzy Jeśli odległości obliczane są pomiędzy obiektami, to macierz D ma wymiary k×k, gdy zaś między cechami — n×n. Przy obliczaniu odległości między cechami (kolumny) stosuje się te same operacje, co przy odległościach między obiektami (wiersze macierzy), z tym,

że zmieniają się granice sumowania. W takim przypadku np. wzór (8) przyjmuje postać:

$$d_{rs} = \left[\frac{1}{k} \sum_{j=1}^{k} (x'_{jr} - x'_{js})^2 \right]^{1/2} \qquad (r, s = 1,2, \dots, n) \qquad (12)$$

Obliczanie odległości pomiędzy parami zbioru Ω za pomocą wzorów (7)—(10) opierało się na założeniu, że każda ze zmiennych (cech) określająca jeden z wymiarów przestrzeni klasyfikacji posiada identyczną wagę. Wydaje się, że należałoby uwzględnić w obliczaniu odległości możliwość ważenia obserwacji. Problem ustalenia właściwej funkcji wagowej jest dość skomplikowany i w znacznej mierze powinien opierać się na przesłankach heurystycznych.[5]

Przy konstrukcji funkcji podobieństwa w oparciu o współczynniki podobieństwa wykorzystuje się rachunek korelacyjny. Zależności istniejące między zmiennymi charakteryzują współczynniki korelacji, tworzą macierz R o postaci:

$$R = \frac{1}{K}(Z^T Z) = \begin{bmatrix} 1 & r_{12} \dots & r_{1m} \\ r_{21} & 1 \dots & r_{2m} \\ - & - - - & - \\ r_{m1} & r_{m2} \dots & 1 \end{bmatrix} \qquad (13)$$

gdzie: Z — macierz znormalizowanych wartości cech o wymiarze $n \times m$,

Z^T — macierz transponowana w stosunku do macierzy Z o wymiarach $m \times n$;

n — liczba obserwacji.

Macierz korelacji R spełnia taką samą rolę, jak macierz odległości D.

Jako miarę podobieństwa wykorzystuje się również współczynnik korelacji w ujęciu wektorowym. W takim przypadku miarę podobieństwa dwóch obiektów jest stopień zbieżności przebiegu wektorów (współczyn-

---

[5] Pewne uwagi dotyczące tego zagadnienia znaleźć można w pracach: J. Liczkowski: *Badanie intensywności rolnictwa w ujęciu przestrzennym*, „Postępy Nauk Rolniczych" 1961, nr 6; J. Mikiewicz: *Zagadnienie wyboru cech przy użyciu metod taksonomii wrocławskiej.* Referat na konferencję naukową Polskiego Towarzystwa Biometrycznego, Wrocław 1967.

nik korelacji), czyli *cosinus* kąta pomiędzy wektorami. Współczynnik ten jest określony następująco: [6]

$$\cos W_i W_l = \frac{W_i W_l}{|W_i|\ |W_l|} \tag{14}$$

przy czym:

$$W_i W_l = \sum_{j=1}^{n} x_{ij} \cdot x_{lj}$$

$$|W_i|\ |W_l| = \sqrt{\sum_{j=1}^{n} x_{ij}^2 \cdot \sum_{j=1}^{n} x_{lj}^2} \tag{15}$$

gdzie: $W_i W_l$ — iloczyn i-tego i l-tego wektora,
$|W_i|\,|W_l|$ — iloczyn skalarny i-tego i l-tego wektora,
x_{ij} — wartość j-tej zmiennej dla i-tego obiektu,
x_{lj} — wartość j-tej zmiennej dla l-tego obiektu.

Dla danych binarnych współczynniki podobieństwa oblicza się z tzw. dwójkowej tablicy podobieństwa badanych obiektów, która ma postać:

Obiekt i-ty

Obiekt l-ty	a	b	a+b
	c	d	c+d
	a+c	b+d	a+b+c+d

gdzie: a — liczba cech występujących równocześnie w i-tym i l-tym obiekcie,
b — liczba cech występujących w obiekcie i-tym a nie występujących w obiekcie l-tym,
c — liczba cech występujących w obiekcie j-tym a nie występujących w obiekcie l-tym,
d — liczba cech nie występujących w obiekcie i-tym i l-tym.

Współczynniki podobieństwa z dwójkowej tablicy podobieństwa obliczane są jako kombinacje elementów tej tablicy. Przykładowo można je obliczyć następująco:

$$W_p = \frac{a+d}{ad+bc} \tag{16}$$

---

[6] J. J. Parysek, L. Wojtasiewicz: *Metody analizy regionalnej i metody planowania regionalnego*. PAN, KPZK, Studia tom LXIX, Warszawa 1979, s. 69.

$$W = \frac{ad}{\sqrt{(a+c)\quad(a+b)\quad(b+d)\quad(c+d)}} \tag{17}$$

Należy zwrócić uwagę na fakt, że interpretacja odległości i współczynników podobieństwa jest odmienna. Rosnąca wartość odległości wskazuje na brak podobieństwa obiektów, których ta odległość dotyczy. Natomiast wzrost współczynnika podobieństwa świadczy o podobieństwie badanych obiektów.

W każdym zadaniu klasyfikacji można wyróżnić następujące elementy składowe:

1) ustalenie zbioru klas (alfabetu klas, obrazów) S. Jeśli zbiór ten jest skończony, to jego elementami są S_i $(i = 1, 2, ..., M)$;

2) dokonanie wyboru własności obiektów, czyli cech charakteryzujących pojedynczą realizację obrazu (realizacja obrazu — to każdy obiekt zbioru reprezentujący dany obraz). Oznaczmy zbiór tych cech przez X, a jego elementy przez X_j $(j = 1, 2, ..., n)$;

3) przyjęcie określonego kryterium klasyfikacyjnego, czyli zasady, według której należy podejmować decyzje, do jakiego obrazu zaliczyć rozpoznawany obiekt. Oznaczmy tę zasadę decyzyjną przez D, a zbiór wartości funkcji decyzyjnej, przy których dany obiekt należy zaliczyć do i-tego obrazu przez D_i;

4) ustalenie wielkości strat spowodowanych błędami klasyfikacji, czyli ustalenie efektywności klasyfikacji (ściślej: ustalenie sposobu pomiaru strat oraz określenie ich poziomu). Oznaczmy wielkość tych strat symbolem E.

W zależności od wstępnych informacji o S, X, D i E (lub przyjętych założeń o tych zbiorach i wielkościach) można wyróżnić cztery elementarne zadania klasyfikacji.

Pierwszym zadaniem jest wybór kryterium klasyfikacji pozwalającego podzielić elementy zbioru Ω scharakteryzowane przy pomocy zbioru cech X pomiędzy klasy S_i (zadane z góry), ponosząc przy tym straty nie większe od E. Zadanie to można w skrócie zapisać następująco:

$$[D/S, X, E]$$

gdzie symbol występujący przed kreską oznacza nieokreślony człon zadania, podczas gdy pozostałe elementy składowe (po kresce) są znane *ex ante*. Zadanie tego typu może również polegać na porządkowaniu nowo pojawiającego się obiektu (nie będącego elementem wyjściowego zbioru Ω) do odpowiedniej klasy S_i. W takim przypadku mówimy o zadaniu klasyfikacji z nauczycielem (lub uczeniem z nagradzaniem). Nazwa „klasyfikacja z nauczycielem wywodzi się stąd, że teoria rozpoznawania

obrazów zajmuje się działaniem dwóch układów: człowieka (nauczyciela) i maszyny (ucznia) [7]. Uczenie z nauczycielem polega na takim współdziałaniu tych układów, że nauczyciel demonstruje uczniowi obiekty, a ten przydziela je do odpowiednich klas. Do rozwiązywania zadań typu pierwszego wykorzystuje się metody klasycznej analizy dyskryminacyjnej, gdyż zagadnienie to można sformułować następująco: dane są wielowymiarowa zmienna losowa X, zmienna losowa Y realizująca wartości równe numerom poszczególnych klas, znane są rozkłady warunkowe typu $F(x/y_i)$ $i = 1, 2, ..., M$, rozkład zmiennej Y oraz macierz stopnia strat stopnia M. W tych warunkach należy podać regułę decyzyjną minimalizującą np. przeciętne straty błędnej klasyfikacji (straty przy bezbłędnej klasyfikacji wynoszą zero).

Drugi typ zadania można określić mianem redukcji wymiarów przestrzeni lub minimalizacji opisu. Rozwiązaniem zadania jest wskazanie takiego podzbioru X (zbiór cech), który pozwoli przyporządkować elementy składowe zbioru Ω (obiekty) do klas (obrazów) S_i przy pomocy kryterium D z minimalnymi stratami E, czyli:

$$[X/S, D, E]$$

Zadanie tego typu należy rozwiązać w ten sposób, by straty spowodowane zmianą ilości informacji (redukcją liczby cech) były jak najmniejsze w sensie funkcji E. Jak łatwo zauważyć tego typu zagadnienie jest identyczne z problemem doboru zmiennych objaśniających do modeli ekonometrycznych.[8]

Trzeci typ zadania klasyfikacji można zapisać następująco:

$$[S/X, D, E]$$

W zadaniu tym chodzi więc o podział elementów zbioru Ω opisanych przy pomocy zbioru własności (cech) X na klasy S_i posługując się przy tym kryterium decyzyjnym D przy zachowaniu efektywności klasyfikacji na poziomie E. Zadania tego typu określane są mianem taksonomii, automatycznej klasyfikacji, grupowania (cluster analysis), samouczeniem (uczeniem bez nauczyciela).

Należy zwrócić uwagę na formalne podobieństwo zadań typu drugiego i trzeciego. W obu typach zadań należy bowiem dokonać określonego

---

[7] B. B. Rozin: *Teoria rozpoznawania obrazów w badaniach ekonomicznych.* Warszawa 1979, s. 11.

[8] Dla modeli ekonometrycznych z dyskretną zmienną objaśniającą analogia jest zupełna, natomiast w przypadku modeli ze zmiennymi ciągłymi należy założyć, że zbiór alfabetu klas jest mocy *continuum.*

grupowania (redukcji przestrzeni). Jednakże w zadaniu typu drugiego redukcja odbywa się w przestrzeni cech, natomiast w zadaniu typu trzeciego — w przestrzeni obiektów.

Czwarty typ zadania można określić mianem optymalizacji. W zadaniu tym należy określić poziom strat E ponoszonych w trakcie procesu klasyfikacji elementów zbioru Ω o własnościach X pomiędzy klasy S_i w oparciu o kryterium D, czyli:

$$[E/S, \overset{*}{X}, D]$$

Dodać należy, że zadania tego typu stanowią zazwyczaj uzupełnienie poprzednich typów, a nie stanowią odrębnego zadania. I tak jeśli np. w problemie klasyfikacji poziom strat dotyczy błędnego zaklasyfikowania pojedynczej realizacji (straty jednostkowe), to rozwiązując zadanie typu czwartego (już po zakończeniu klasyfikacji) jesteśmy w stanie podać poziom strat przeciętnych.

Powyższe cztery zadania klasyfikacyjne zostały określone mianem zadań elementarnych (prostych). W praktyce badań statystycznych nie zawsze dysponujemy, tak dużą ilością informacji wyjściowych (trzy spośród czterech członów muszą być znane *a priori*). Dlatego też zachodzi konieczność rozwiązywania tzw. zadań kombinowanych. W tego typu zadaniach mogą być znane dwie (lub nawet jedna) składowe procedur klasyfikacyjnych. Przykładowo można wymienić takie zadania, jak: [X, D/S, E], [S, X/D, E], [E/S, X, D] itp. Mogą również wystąpić sytuacje, w których żaden element procedury klasyfikacyjnej nie jest znany *ex ante*. Mówimy wówczas o zadaniach złożonych. Zasadnicze znaczenie posiadają jednak zadania elementarne, gdyż zarówno kombinowane, jak i złożone można sprowadzić do zadań prostych. Przykładowo zadanie kombinowane w postaci [S, D/X, E] można rozwiązać rozpatrując szereg zadań elementarnych [S/X, D, E] zakładając różne możliwe kryteria klasyfikacyjne. Należy się jednak wtedy liczyć ze znacznym wzrostem pracochłonności, nawet przy zastosowaniu maszyn cyfrowych.

РЕЗЮМЕ

В статье приводится общая характеристика исследовательских проблем охватываемых термином „многоизмерительный статистический анализ". В частности, дано определение термина „классификация", виды классификаций и способы построения функций сходства, позволяющих зачислять отдельные объекты к определенным гомогенным классам. В конечной части статьи определены четыре основные (элементарные) задачи классификации.

SUMMARY

In the article an analysis was carried out as to the characterization of research problems included in so-called multi-dimensional statistical analysis (WAS). In particular, a definition of the term "classification" was provided, there were also given the types of classifications and the methods of constructing the functions of similarity which make it possible to group definite objects under homogeneous classes. The last part of the article specified the four fundamental (elementary) tasks of classifications.